Data Analysis Assignment 2

Viraj Karambelkar
150260003

Hrishikesh T Iyer
150260024

Harikrishnan KP
150260026

Nitin Srirang
150260027

October 13, 2016

Contents

1 Report for Assignment 2

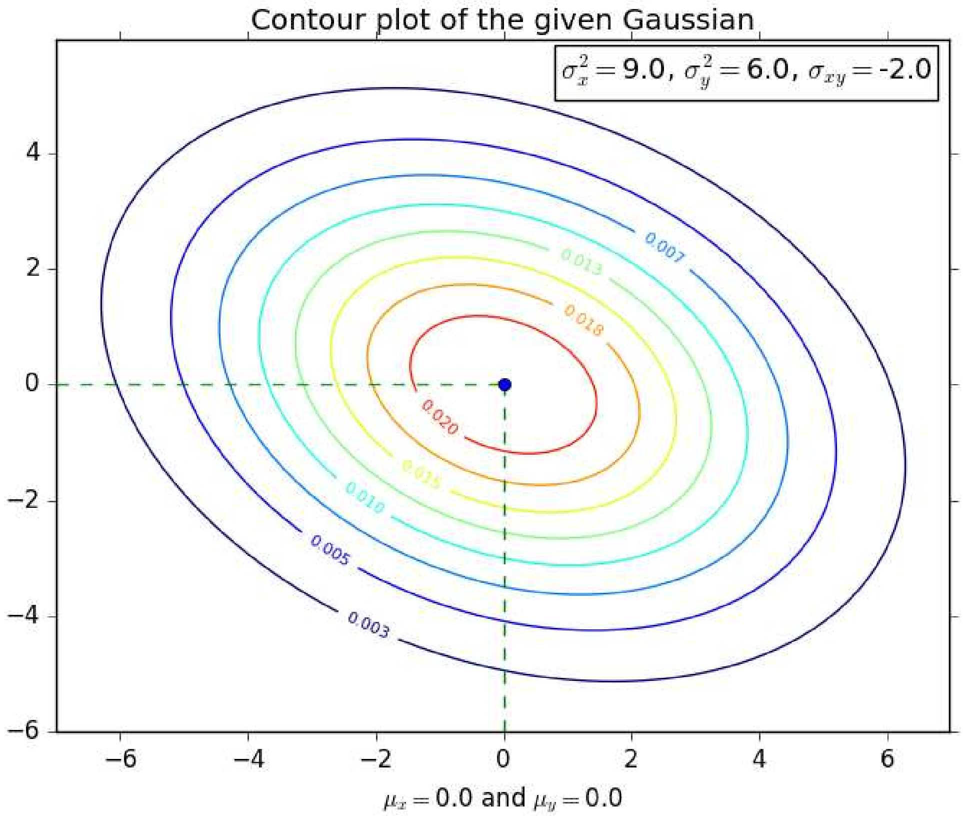
1.1 Multivariate Gaussian Distribution

The problem involves 2 correlated Gaussian random variables x and y with 0 mean.
Their covariance matrix is given by :

$$C = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 9 & -2 \\ -2 & 6 \end{bmatrix}$$

1.1.1 Contour Plot

The contour plot obtained for the probability distribution function of x and y is :



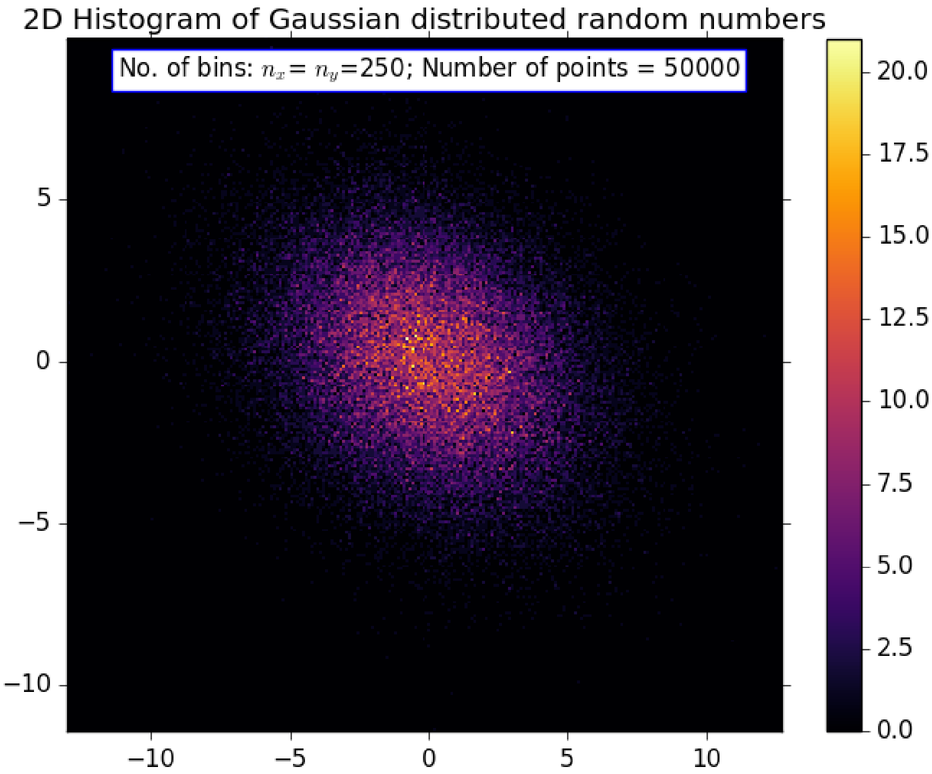
The contours are ellipses with their major axes inclined to the positive X axis at an obtuse angle beacause of the negative correlation between the two random variables.
The eigenvalue equation for matrix C is $(6-\lambda)(9-\lambda)-4=0$
The eigenvalues are 5 and 10 for which the corresponding eigenvectors can be found.
After normalization the eigenvectors are :

$$x' = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad y' = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Hence the major axes of the ellipses are along $x+2y=0$ and the minor axes are along $2x-y=0$.

1.1.2 2D Histogram of random pairs generated from the distribution

50000 pairs of random numbers (x,y) having this distribution were generated and a 2D histogram was plotted.

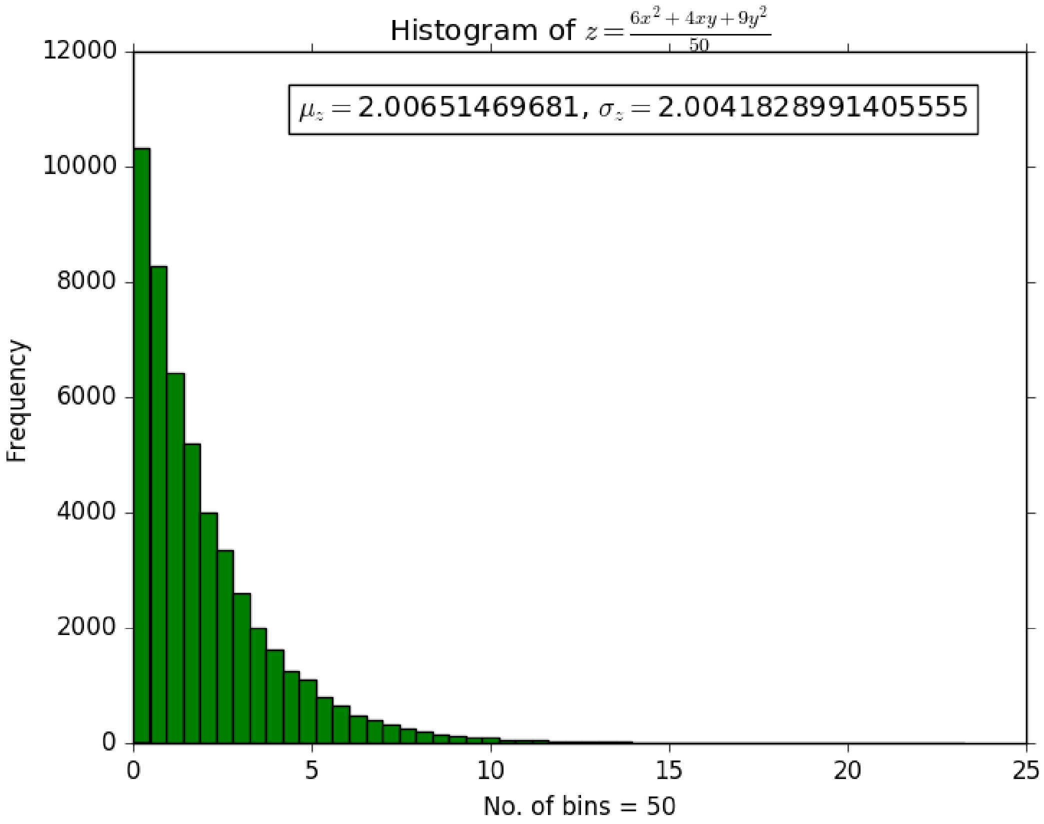


250 bins were cosidered for plotting the histogram. The height of the histogram is shown using the color scheme 'Inferno'.

The maximum number of points in the 2D histogram are towards the center (0,0) and decreases as we move outward. The height of the histogram at various points is clearly in agreement with the contour diagram.

1.1.3 Distribution of z

A new variable z is generated for each of the 50000 pairs of x and y that are generated.
$z=\frac{1}{50}(6x^2+4xy+9y^2)$
A histogram is plotted for the values of z.

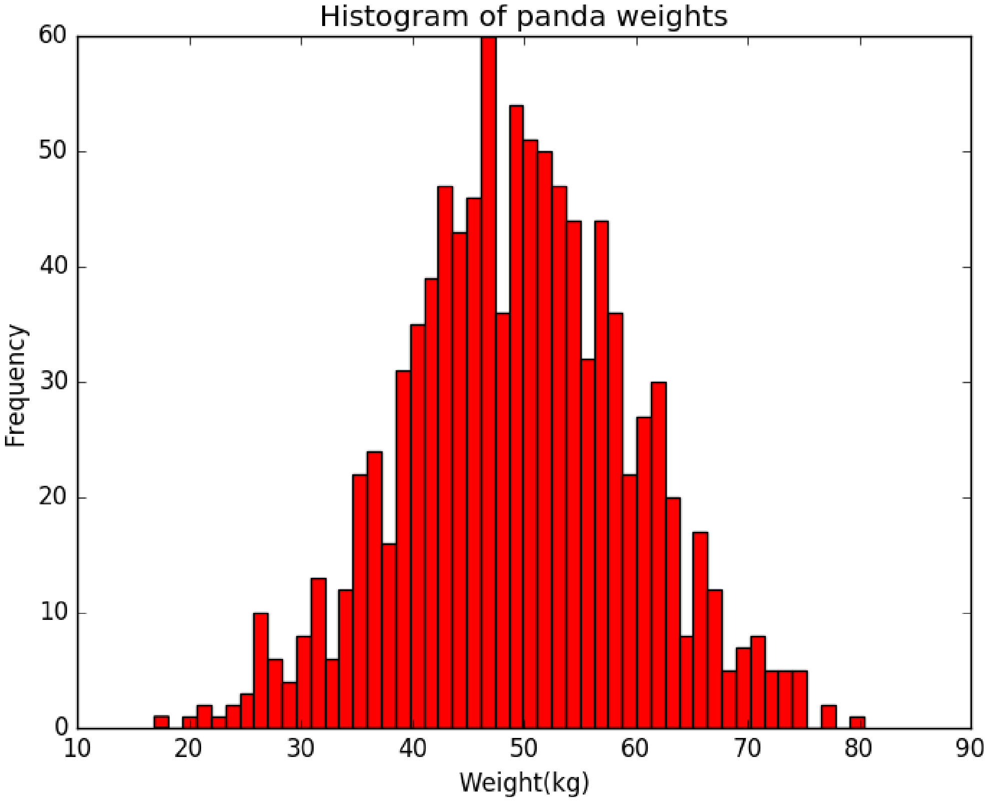


The new variables x’ and y’ have a covariance matrix : $C' = \begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix}$ Let a new variable w be given by $w = \frac{x'^2}{10} + \frac{y'^2}{5}$ w is a sum of squares of Gaussian distributed variables divided by the sum of their respective variances and hence is chi-squared distributed. On substituting values of x’ and y’(i.e in terms of x and y), w simplifies to z. Hence z has a χ^2 distribution which is confirmed by the histogram.
The mean value of z is 2.00651469681 with a standard deviation of 2.00418289914.

1.2 Panda Statistics

The weight of 100 baby pandas is read from the file *"pandas.txt"* and stored in an array.

- The mean weight of the pandas is calculated to be **49.41660999999994** with an error(standard deviation) of **10.210954715603448**.
- The size of typical fluctuations of the weight of each baby panda about the mean is calculated by summing up the modulus value of the difference of each individual weight from the sample mean and dividing it by the number of pandas under consideration. This value works out to be **8.110777120000005**.

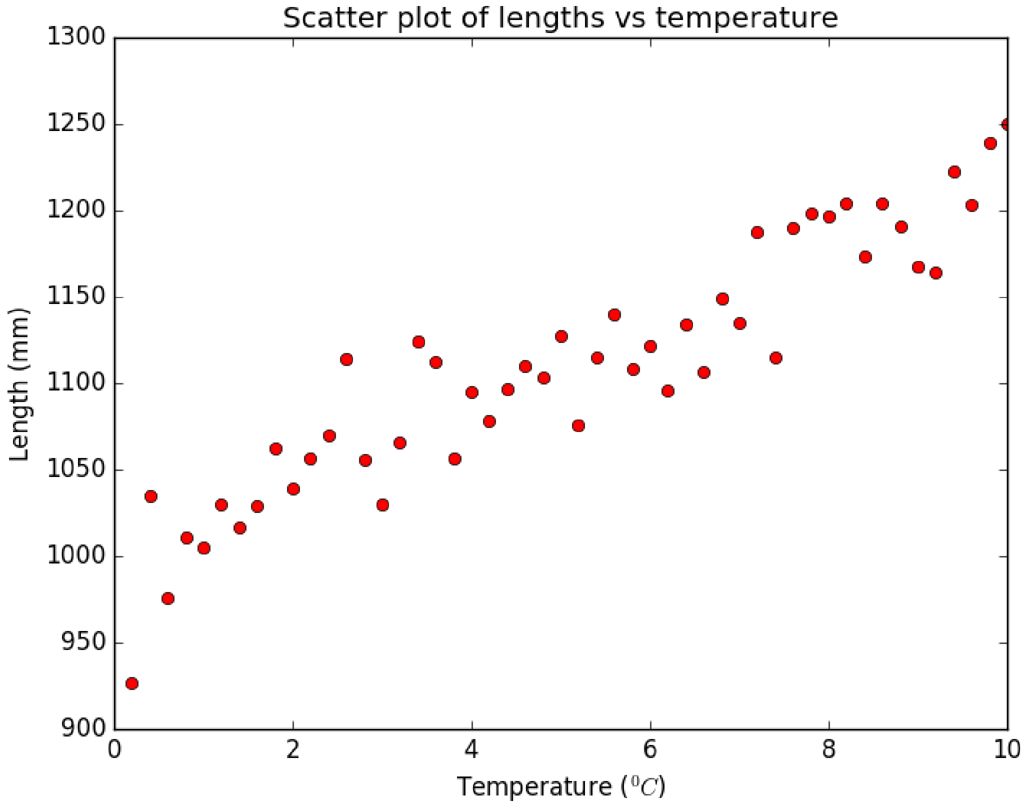


1.3 Linear Expansion of Aluminium Rod

The readings of the length of an aluminium rod as a function of temperature is read from the file *"linearexpansion.csv"*.

1.3.1 Scatter Plot

The scatter plot obtained for the given data is:



1.3.2 Best Fit Line

Let us consider y=mx+c as the equation for the best-fit line. Analytically the value of m and c can be found:
$SS = \sum_{i=0}^{n}(y_i - (mx_i + c))^2$

where SS is the sum of squares.

$\frac{\partial SS}{\partial m} = 0 \Rightarrow \sum_{i=0}^{n} x_i(y_i - (mx_i + c)) = 0$

$\frac{\partial SS}{\partial c} = 0 \Rightarrow \sum_{i=0}^{n}(y_i - (mx_i + c)) = 0$

$\sum_{i=0}^{n} y_i = m\sum_{i=0}^{n} x_i + Nc$

$\Rightarrow c = \overline{y} - m\overline{x}$

$\sum_{i=0}^{n} x_i y_i = m\sum_{i=0}^{n} x_i^2 + Nc\overline{x}$

$\sum_{i=0}^{n} x_i y_i = m\sum_{i=0}^{n} x_i^2 + N\overline{x}(\overline{y} - m\overline{x})$

$m = \frac{\sum_{i=0}^{n} x_i y_i - \overline{x}\sum_{i=0}^{n} y_i}{\sum_{i=0}^{n} x_i^2 - N\overline{x}^2}$

Let σ be the single measurement error for each y
$\sigma_m^2 = \frac{\sum_{i=0}^{n}((x_i - \overline{x})^2 \sigma_{yi}^2}{\sum_{i=0}^{n} x_i^2 - N\overline{x}^{2^2}}$

$\sigma_m^2 = \frac{\sigma^2}{(\sum_{i=0}^{n} x_i^2) - N\overline{x}^2}$

$c = \frac{\sum_{i=0}^{n} y_i}{N} - m\overline{x}$

$\sigma_c^2 = \frac{1}{N^2}\sum_{i=0}^{n}(\sigma_{yi}^2 + \overline{x}^2\sigma_m^2$

$\sigma_c^2 = \frac{\sigma^2}{N} + \overline{x}^2\sigma_m^2$

Let x_0 be the temparature at extrapolated point and y_0 be the extrapolated length. Then
$\sigma_{y_0}^2 = x_0^2\sigma_m^2 + \sigma_c^2\sigma_{y_0}^2 = x_0^2\left[\frac{\sigma^2}{(\sum_{i=0}^{n} x_i^2) - N\overline{x}^2}\right] + \frac{\sigma^2}{N} + \overline{x}^2\left[\frac{\sigma^2}{(\sum_{i=0}^{n} x_i^2) - N\overline{x}^2}\right]$

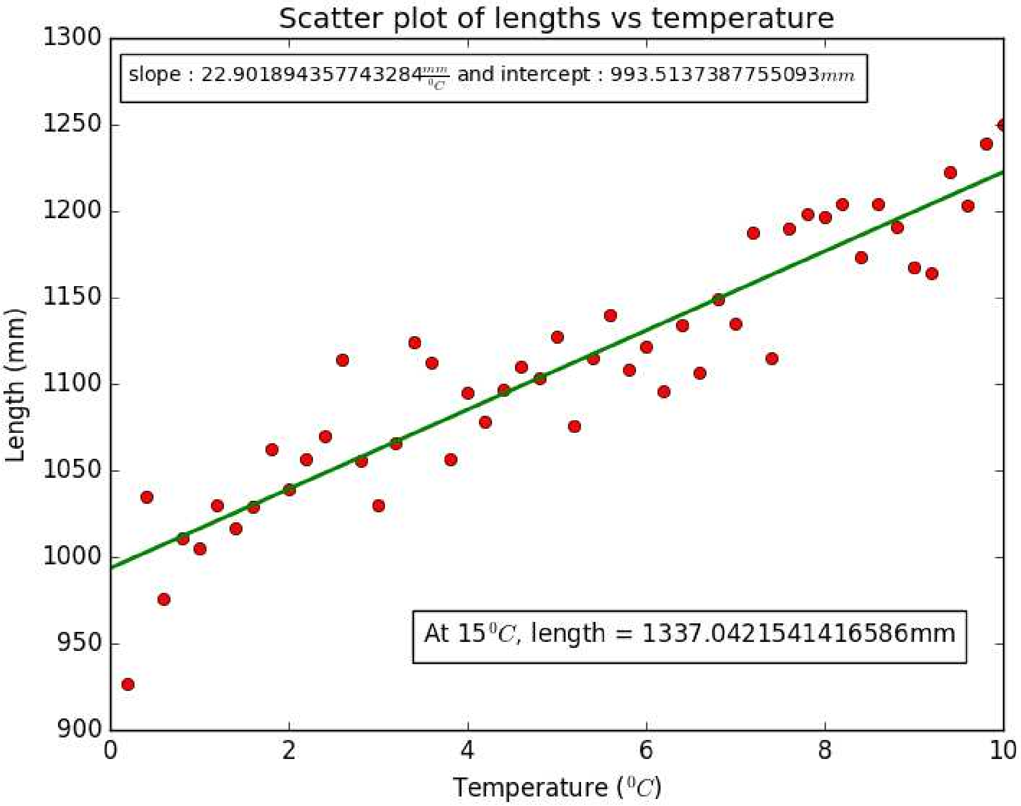
$\sigma_{y_0}^2 = \sigma^2\left[\frac{1}{N} + \frac{(\overline{x}^2 + x_0^2)}{(\sum_{i=0}^{n} x_i^2) - N\overline{x}^2}\right]$

$where \quad \sigma^2 = \frac{1}{N-2}\sum_{i=0}^{n}(y_i - \hat{y})^2$

where $\hat{y} = mx_i + c$.

- The value of m obtained is *22.901894357743284* and value of c is *993.5137387755093*.
- The error in each single measurement is *72.55844732036144 mm*.
- Linear extrapolation upto 15 degree Celsius gives an expected length of *1337.0421541416586 mm*.
- Error in length at 15 degree Celsius is *57.25523889170834 mm*.

The best fit line obtained has been plotted over the scatter diagram.



2 Team Responsibilities

The roles for the group members for this assignment are as follows:

Nitin Srirang	Group Leader
Hrishikesh T Iyer	Coder
Viraj Karambelkar	Web Manager
Harikrishnan KP	Report Writer

3 Website

The link to our website is The WIMPy Kids-Bringing Numbers to Life. The code and report for each week's assignment can be found on this site. The code has also been uploaded on Github, the link for the assignment repository is Data Analysis Assignment Repo. The code for this week's assignment is in this branch of the repository.